Protein loops, solitons and side-chain visualization with applications to the left-handed helix region

Martin Lundgren,[1, *] Antti J. Niemi,[2, 1, †] and Fan Sha[1, ‡]

[1] *Department of Physics and Astronomy, Uppsala University, P.O. Box 803, S-75108, Uppsala, Sweden*
[2] *Laboratoire de Mathematiques et Physique Theorique CNRS UMR 6083,*
Fédération Denis Poisson, Université de Tours, Parc de Grandmont, F37200, Tours, France

Folded proteins have a modular assembly. They are constructed from regular secondary structures like α-helices and β-strands that are joined together by loops. Here we develop a visualization technique that is adapted to describe this modular structure. In complement to the widely employed Ramachandran plot that is based on toroidal geometry, our approach utilizes the geometry of a two-sphere. Unlike the more conventional approaches that only describe a given peptide unit, ours is capable of describing the entire backbone environment including the neighboring peptide units. It maps the positions of each atom to the surface of the two-sphere exactly how these atoms are seen by an observer who is located at the position of the central C_α atom. At each level of side-chain atoms we observe a strong correlation between the positioning of the atom and the underlying local secondary structure with very little if any variation between the different amino acids. As a concrete example we analyze the left-handed helix region of non-glycyl amino acids. This region corresponds to an isolated and highly localized residue independent sector in the direction of the C_β carbons on the two-sphere. We show that the residue independent localization extends to C_γ and C_δ carbons, and to side-chain oxygen and nitrogen atoms in the case of asparagine and aspartic acid. When we extend the analysis to the side-chain atoms of the neighboring residues, we observe that left-handed β-turns display a regular and largely amino acid independent structure that can extend to seven consecutive residues. This collective pattern is due to the presence of a backbone soliton. We show how one can use our visualization techniques to analyze and classify the different solitons in terms of selection rules that we describe in detail.

I. INTRODUCTION

The Ramachandran plot [1], [2] is the paradigm technique of protein visualization. It describes backbone atoms in a peptide group around a given C_α carbon in terms of dihedral rotations. Ramachandran plot can also been extended to the side-chain atoms in terms of the dihedral rotamers. This gives rise to Janin plot [3] and its variants. In the present article we develop new visualization techniques to describe proteins. Our goal is to visualize all atoms both in a given peptide unit and those in the neighboring units, beyond the regime of the Ramachandran plot. This will enable us to search for new relations between the positioning of various atoms and the backbone geometry. Our approach draws from developments in three dimensional visualization that have taken place after the Ramachandran plot was originally introduced [4], [5]. In particular, in lieu of toroidal geometry we utilize the geometry of a two-sphere. It enables us to describe the various atoms exactly as they are seen by an observer who roller-coasts along the backbone. Of particular interest to us is the visual analysis of the modular components of which all folded proteins are built. These have been recently identified as the soliton solutions to a generalized discrete nonlinear Schrödinger equation (DNLS) [6]-[8].

Soliton solutions to nonlinear difference equations share a long history with biological physics of proteins. The discrete version of the nonlinear Schrödinger equation is an embodiment of this relationship. It was originally introduced by Davydov [9] to describe energy transfer along the protein α-helices. Subsequently the DNLS equation has found many additional applications in biological physics and elsewhere [10]. The DNLS equation has also the remarkable mathematical property of integrability, it is commonly viewed as the archetype integrable system [11].

When the DNLS soliton propagates along the α-helix, the protein changes its shape. In [6], [7] it has been shown that when the soliton becomes trapped, the protein folds. It now appears that practically all folded proteins can be built in a modular fashion from a relatively small number of such trapped solitons [8]. In the present article we combine the notion of soliton with modern visualization techniques [12]. We are particularly interested in the ramifications of the backbone DNLS soliton in protein side-chain geometry. Our ultimate goal is to develop a graphical characterization and eventually a full classification of protein structures in terms of their soliton modules. As a prelude, we here utilize the soliton concept to visually inspect and analyze those protein conformations that are located in the left handed α-helix (L-α) region of the Ramachandran plot [1], [2]. This region is a relatively small subset of all different protein conformations, and as such amenable to an explicit analysis.

Of particular interest to us are the common geometric aspects of the asparagine (ASN) and aspartic acid

[*] Martin.Lundgren@physics.uu.se
[†] Antti.Niemi@physics.uu.se
[‡] fansha0559@gmail.com

(ASP). Asparagine is the predominant residue in the so-called non-glycyl L-α region. According to the prevailing point of view this is due to a *localized* non-covalent attractive carbonyl-carbonyl interaction between the side-chain and backbone [13]-[15]. Such a carbonyl-carbonyl interaction can only be present in ASN, ASP, glutamine (GLN) and glutamic acid (GLU). Indeed, the propensity of ASP that is structurally very similar to ASN is also clearly amplified in the L-α region, while the somewhat lower propensity of GLN and GLU has been explained in the literature to be a consequence of steric suppressions [13].

Here we show that the presence of a L-α site goes beyond the regime of a single peptide unit. We find that it involves a coordinated interplay of up to seven consecutive amino acids. We argue that this extended correlation over several amino acids is symptomatic to solitons. We perform a detailed visual investigation and propose a graphical classification of these solitons. We argue that all protein structures could be characterized and classified similarly, in terms of general selection rules that we formulate. We find that the continuous geometry of the two-sphere gives a more perceptible characterization of protein conformations than the toroidal Ramachandran plot. In fact, the three dimensional visualization techniques we utilize have been largely introduced and developed after the publication of [1], [2]. Our approach exploits the properties of a piecewise linear *framed* chain, as it is being applied to visualization problems in aircraft and robot kinematics, stereo reconstruction, and increasingly in computer graphics and virtual reality [4], [5]. In these applications different framings correspond to different camera gaze positions, that one introduces and varies for the purpose of extracting diverse and complementary information on geometrical aspects and physical properties of the system under investigation. However, largely due to the success and systematics provided by Ramachandran plot, thus far this kind of approach has been sparsely applied to the analysis of protein conformations. Among our goals is to demonstrate that these modern visualization techniques can provide a powerful complementary tool for the visual description of folded proteins. In particular, they enable the study of visual correlations between nearby peptide units, which is not possible in the Ramachandran approach that is limited to to describe a single peptide unit only.

Finally, we note that the investigation of the physical properties of our concrete examples ASN and ASP is also of substantial biological interest. These two amino acids are more frequently than any other amino acid subject to *in vivo* post-translational modifications including spontaneous nonenzymatic deamidation from ASN to ASP [16] and racemization from L-ASP into D-ASP [17]. These processes are presumed to have consequences to cellular and organismal ageing [16], [18]. They might also have a rôle in enhancing the emergence of amyloid based neurodegenerative diseases [18], [19].

II. FRAMING

We interpret a protein backbone in terms of framed chain, with vertices located at the C_α carbons [12]. Depending on the application, the framing can be introduced in various different ways. Examples include the geometric Frenet frame [4], [5], the geodesic Bishop frame [20], and protein specific C_β carbon frame that we obtain by utilizing the direction of the C$_\beta$ carbon along a protein backbone to construct an orthonormal framing [12]. Here we propose that in particular the Frenet framing provides a powerful tool for protein side-chain visualization, also beyond our explicit example of the L-α Ramachandran region. The additional advantage of the Frenet framing is that it relates directly to an energy function. But we also advertise the closely related C$_\beta$ framing that may sometimes have certain visual advantages.

The framing of a piecewise linear chain is conventionally based on the Denavit-Hartenberg [21] formalism. This formalism was originally introduced in robotics but has been subsequently extensively applied also in other disciplines. Here we resort to a variant, that has been developed in [12] for the purpose of framing protein backbones. It utilizes the transfer matrix formalism [11] to describe a protein with N residues using the coordinates $\mathbf{r}_i$ of the backbone C_α carbons ($i = 1, ..., N$). These coordinates can be downloaded from the Protein Data Bank (PDB) [22]. For each of the segments that connect the backbone C_α central carbons we compute the unit length tangent vector $\mathbf{t}_i$, binormal vector $\mathbf{b}_i$ and normal vector $\mathbf{n}_i$ using

$$\mathbf{t}_i = \frac{\mathbf{r}_{i+1} - \mathbf{r}_i}{|\mathbf{r}_{i+1} - \mathbf{r}_i|}$$

$$\mathbf{b}_i = \frac{\mathbf{t}_{i-1} \times \mathbf{t}_i}{|\mathbf{t}_{i-1} \times \mathbf{t}_i|} \tag{1}$$

$$\mathbf{n}_i = \mathbf{b}_i \times \mathbf{t}_i$$

Thus the tangent vector $\mathbf{t}_i$ points from the i^{th} central C_α carbon to the direction of the $(i+1)^{th}$ central C_α carbon, the way how it is seen by an observer who is located at the position of the i^{th} carbon. The $\mathbf{b}_i$ and $\mathbf{n}_i$ determine a frame that enables the observer to orient herself at the location $\mathbf{r}_i$, on the plane that is orthogonal to the direction $\mathbf{t}_i$. Together the right-handed triplet $(\mathbf{n}_i, \mathbf{b}_i, \mathbf{t}_i)$ constitutes the orthonormal discrete Frenet frame for each residue along the backbone chain, with base at the position of the vertex $\mathbf{r}_i$. The corresponding backbone bond $\kappa_{i+1,i} \equiv \kappa_i$ and torsion $\tau_{i+1,i} \equiv \tau_i$ angles can be computed from (1) as follows,

$$\cos\kappa_i \;=\; \mathbf{t}_{i+1} \cdot \mathbf{t}_i \tag{2}$$

$$\cos\tau_i \;=\; \mathbf{b}_{i+1} \cdot \mathbf{b}_i \tag{3}$$

Alternatively, if the bond and torsion angles are known we can construct the frames iteratively by starting from the N terminus and using [12]

$$\begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_{i+1} = \mathcal{R}_{i+1,i} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_{i}$$

$$= \exp\{\kappa_i T^2\} \cdot \exp\{\tau_i T^3\} \begin{pmatrix} \mathbf{n} \\ \mathbf{b} \\ \mathbf{t} \end{pmatrix}_{i} \qquad (4)$$

where the T^a $(a = 1, 2, 3)$ are the adjoint SO(3) Lie algebra generators. Once the $\mathbf{t}_i$ have been constructed from (4) and the bond lengths $s_i = |\mathbf{r}_{i+1} - \mathbf{r}_i|$ have been determined we recover the entire backbone from

$$\mathbf{r}_k = \sum_{i=0}^{k-1} s_i \cdot \mathbf{t}_i \qquad (5)$$

The set of all Frenet frames defines a framing of the backbone. According to (2)-(4) the bond and torsion angles are link variables, they relate a frame at the vertex $\mathbf{r}_i$ to a frame at the vertex $\mathbf{r}_{i+1}$. We note that the definition of the bond angle involves three vertices while the definition of the torsion angle involves a total of four vertices.

The C_β framing [12] is a complement to the Frenet framing. It can be introduced for all non-glycyl residues. We define these frames similarly, in terms of three mutually orthogonal unit vectors at each C_α carbon. Consequently the Frenet framing and the C_β framing are related to each other by peptide unit dependent SO(3) rotations. The first unit vector of the C_β basis is obtained as follows,

$$\mathbf{s}_i = \frac{\mathbf{r}_{\beta,i} - \mathbf{r}_{\alpha,i}}{|\mathbf{r}_{\beta,i} - \mathbf{r}_{\alpha,i}|}$$

Here $\mathbf{r}_{\alpha,i}$ is the location of the ith C_α atom, and $\mathbf{r}_{\beta,i}$ is the location of the corresponding C_β atom. The second unit vector is

$$\mathbf{p}_i = \frac{\mathbf{s}_i \times \mathbf{t}_i}{|\mathbf{s}_i \times \mathbf{t}_i|}$$

where $\mathbf{t}_i$ is the Frenet frame unit tangent vector. Finally, the third unit vector in the C_β frame is

$$\mathbf{q}_i = \mathbf{s}_i \times \mathbf{p}_i$$

Since $(\mathbf{s}_i, \mathbf{p}_i, \mathbf{q}_i)$ is an orthonormal frame located at each C_α, it can be used like the Frenet frame to visualize the various atoms along the protein backbone. Moreover, since

$$\mathbf{t}_i = \mathbf{p}_i \times \mathbf{s}_i$$

we can likewise use the C_β framing to construct the entire backbone using (5).

We wish to employ the various frames together with the discrete Frenet equation (4) to inspect the structure of folded proteins. As our principal data set we utilize all those proteins that are presently in PDB and have an overall resolution that is better than 2.0 Ångström. We introduce no additional curation or data pruning in this set. But we have confirmed that all our results and conclusions stand when we restrict ourselves to those proteins with resolution better than 1.5 Å and with less than 30% homology relation, or to those proteins that have a resolution which is better than 1.0 Å. Finally, as a control set we also utilize the highly curated version v3.3 Library of chopped PDB files for representative CATH domains [23]. Since the conclusions we draw are indifferent of the data set that we use we only describe explicitly the results for the first one, as it allows for the visually most complete presentation.

III. BACKBONE MAPPING

We start by describing how to visualize the protein backbone in terms of the Frenet frames [12]. Here we go beyond the regime of the Ramachandran plot, that does not provide any direct visual correlation between neighboring peptide groups. We introduce an observer who maps all the atoms in the protein by traversing along the backbone. The observer moves between the C_α carbons like on a roller-coaster with an orientation that is determined by the discrete Frenet framing: We take the base of the tangent vector $\mathbf{t}_i$ defined in (1) to be at the location $\mathbf{r}_i$ of the i^{th} central C_α carbon. The tip of $\mathbf{t}_i$ then determines a point on the surface of a unit two-sphere that surrounds our observer at the location of this C_α carbon. The observer uses this two-sphere to constructs a map of the various atoms exactly the way how she sees them on the surface of the sphere, as if the atoms were stars in the sky. For this she always orients the two-sphere at the site i so that the north-pole coincides with the tip of $\mathbf{t}_i$ *i.e.* the north-pole is always in the direction of the next C_α at the site $\mathbf{r}_{i+1}$. She takes the bond angle to measure the latitude of the two-sphere from its north pole. The torsion angle measures the longitude starting from the great circle that passes both through the north pole and through the tip of the binormal vector $\mathbf{b}_i$. In terms of these angles she can characterize the direction of the vector $\mathbf{t}_{i+1}$ *i.e.* the direction towards site $\mathbf{r}_{i+2}$ to which the roller coaster turns at the next C_α carbon. Consequently she acquires information about the geometric relations between neighboring peptide units, and this goes beyond the regime of the Ramachandran plot. She proceeds as follows:

She first translates the center of the two-sphere from the location of the i^{th} central carbon towards its north-pole and all the way to the location of the $(i+1)^{th}$ central carbon, without introducing any rotation of the sphere. She then records the direction of $\mathbf{t}_{i+1}$ as a point on the surface of the two-sphere. This defines the corresponding coordinates (κ_i, τ_i) and marks a point on the map. It gives an instruction to the observer at the point $\mathbf{r}_i$, how she should turn at site $\mathbf{r}_{i+1}$, to reach the $(i+2)^{th}$ central C_α carbon at the point $\mathbf{r}_{i+2}$.

She then continues to construct the mapping with the next C_α carbon along the backbone. She rotates the two-sphere at $\mathbf{r}_{i+1}$ so that the north pole of the rotated sphere coincides with the tip of $\mathbf{t}_{i+1}$, and so that the torsion angle measures the longitude from the great circle determined by the north-pole and the tip of $\mathbf{b}_{i+1}$. She repeat the procedure for all C_α, until she has mapped the entire backbone. We note that for a folded protein the two vectors $\mathbf{t}_i$ and $\mathbf{t}_{i+1}$ are never exactly parallel to each other so there is never any ambiguity due to an inflection point.

When we repeat this mapping procedure for every C_α in all proteins in our data set, we obtain a (κ, τ) distribution that characterizes the overall geometry of protein backbones. This provides non-local information on the backbone geometry that extends over several peptide units. In particular, we now have a map that shows *exactly* how the central carbons are seen by our roller-coasting observer when she gazes at them from her Frenet frame positions along the backbone.

We find that the C_α distribution for all proteins in our data set determines an annulus on the surface of the two-sphere. For visualization it then becomes convenient to employ the geometry of the stereographically projected two-sphere. It is obtained by projecting our (κ, τ) coordinates to the north pole tangent plane of the two-sphere. If (x, y) are the coordinates of this tangent plane the projection is defined by

$$x + iy = \tan(\frac{\kappa}{2}) \cdot e^{-i\tau} \tag{6}$$

When we perform this projection for all C_α carbons in all proteins that are in our data set and separately display the results for the different groups of α-helices, β-strands, 3/10-helices and loops as these structures are defined in PDB, we arrive at the angular distributions that we show in Figures 1. For our observer who always fixes her gaze position towards the north-pole of the surrounding two-sphere at each C_α carbon, *i.e.* towards the black dot at the center of the annulus, the color intensity reveals the likely direction to which the roller coaster who is located at position $\mathbf{r}_i$ turns at the next C_α carbon, when she starts moving from its location at $\mathbf{r}_{i+1}$ towards $\mathbf{r}_{i+2}$. In particular, the four maps in Figure 1 are in a direct visual correspondence with the way how the Frenet frame observer perceives the backbone geometry.

The four maps in Figure 1 portray non-local features that are not available in conventional Ramachandran plots. Moreover, instead of a discontinuous toroidal square as in the case of the Ramachandran plots, the predominant feature in all of the present maps is that the PDB data is concentrated in a continuous annulus which is roughly between the circles $\kappa_{in} \approx 1$ and $\kappa_{out} \approx \pi/2$. The exterior of the annulus $\kappa > \kappa_{out}$ is an excluded region, it describes conformations that are subject to steric

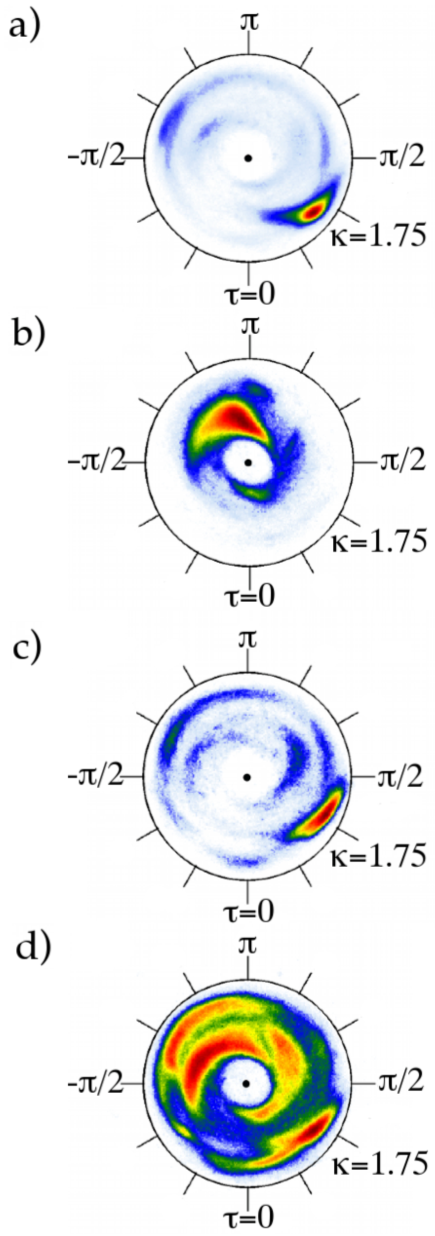

FIG. 1. (Color online:) The four major protein structures: a) α-helices, b) β-strands, c) 3/10-helices and d) loops. We define these structures according to their PDB classification in our 2.0 Å data set. In each Figure the center of the annulus is the north-pole of the two-sphere that surrounds the observer at the position i. This is the direction where the next C_α is located. At this point the bond (latitude) angle $\kappa = 0$. The bond angle measures distance from the center of the annulus so that the south pole where $\kappa = \pi$ corresponds to points at infinity on the plane. The torsion *i.e.* longitude angle $\tau \in [-\pi, \pi]$ increases by 2π when we go around the center of the annulus in counter-clockwise direction. The color coding in all our Figures increases from white to blue to green to yellow to red describes the relative number of conformations in PDB in a log-squared scale. The intensity is proportional to the probability of the direction where the observer turns at the next C_α carbon.

clashes. The interior $\kappa < \kappa_{in}$ is sterically allowed but practically excluded as long as proteins remain in the collapsed phase; The interior region becomes occupied when we cross the Θ-point and proteins assume their unfolded conformations.

We notice that loops appear to have a slightly higher tendency to bend towards left *i.e.* $\tau < 0$. We also note that in the Figures for α, β and 3/10 the blue regions correspond to residues where the present hydrogen-bond based PDB classification is in a miss-match with the geometric structure that is commonly associated with these configurations. Moreover, the Figure reveals that the PDB data displays innuendos of various underlying reflection symmetries: In the Figure 1d (loops) there is a clearly visible mirror of the standard right-handed α-helix region, located in the vicinity of the outer rim with $\kappa \approx 3/2$ and with torsion angle close to the value $\tau \approx -2\pi/3$. A helix in this regime would be left-handed and tighter than the standard α-helix. There is also a clear mirror structure in the Figure 1b for β strands, the standard region is $(\kappa, \tau) \approx (1, \pi)$ and its less populated mirror is located around $(\kappa, \tau) \approx (1, 0)$. The mirror symmetry between the ensuing extended regions persists in the Figure 1d for loops. Finally, in the Figure 1d we observe a small elevated (yellow) region in the vicinity of $(\kappa, \tau) \approx (3/2, -\pi/3)$. This is the region of helices that are spatial left-handed mirror images of the standard α-helices. There is also a (slightly) elevated (green) mirror of this region around $(\kappa, \tau) \approx (3/2, 2\pi/3)$. This is like the $(\kappa, \tau) \approx (3/2\,, -2\pi/3)$ mirror of the standard right-handed α helices.

IV. SIDE-CHAIN MAPPING

We can similarly visualize the geometry of side-chain atoms, as they are seen by our roller-coasting observer. This gives us local information on the given peptide unit. Now the results turn out to be isomorphic to those revealed by the standard Ramachandran plot. Moreover, this enables us to develop a visual complement to the existing rotamer libraries.

We assume that the observer is oriented according to the discrete Frenet framing that is determined by the transfer matrix (4) at each C_α. At the location of the C_α the observer then looks at the side-chain atoms and records the direction of each of them as points on the surface of the two-sphere that surrounds the observer, with the north-pole of the sphere always coinciding with the direction towards the next C_α exactly as in the case of the backbone.

In Figure 2 (top) we display the angular distribution of the C_β carbons on the surface of the two-sphere for all the C_α carbons, as recorded by our Frenet frame observer who is located at the origin of the sphere. Recall that a C_β carbon is present in all non-glycyl residues. We note that our framing is determined entirely in terms of the backbone. According to prevailing paradigm the directions of the C_β carbons should then be directly computable from the geometry of the tetrahedral covalent bond structure of the pertinent C_α carbon. However, Figure 2 (top) reveals that the directions of the C_β carbons are not determined only by the local covalent bond structure. In addition, these directions are clearly subject to secondary structure dependent but amino acid independent nutations. This confirms that at the level of accuracy of our data, the stereochemical restraints fail to be fully universal. They reflect the secondary structure environment [24]-[28]. In fact, despite being based entirely on the C_β atoms the Figure 2 is fully isomorphic to the standard Ramachandran plot, for all amino acids except for glycine that has no C_β.

A important feature of the nutation is the presence of the highly localized, isolated island denoted $\mathtt{L}$-α that is clearly visible in Figure 2 (top). We have confirmed that this isolated island coincides exactly with the conventional non-glycyl $\mathtt{L}$-α region of the standard Ramachandran plot. This is shown in Figure 2 (middle) where we display the direction of the C_β carbons solely for those non-glycyl residues that are in the $\mathtt{L}$-α Ramachandran region. Finally, in the Figure 2 (bottom) we display the discrete Frenet frame distribution of the C_β carbons for those ASN that are located in loops only, according to PDB classification. The relatively high propensity of ASN in the $\mathtt{L}$-α island is prominent.

In the sequel we shall concentrate our attention solely on the isolated $\mathtt{L}$-α island in Figure 2 (middle). We start by noting the propensity of different amino acids in the $\mathtt{L}$-α island. The result (in percent) is shown in Figure 3. This Figure confirms the high propensity of ASN (N) that is also visible from Figure 2 (bottom). We find that ASP (D) has also relatively high relative propensity. But the propensity of histidine (H) is practically equal. Furthermore, several non-carbonylic amino acids have a higher propensity than GLU (E). Finally, the β-branched isoleucine (I), valine (V) and threorine (T) all have clearly suppressed propensities and proline (P) is practically absent, presumably reflecting the presence of steric constraints [13], [15].

We now proceed to map the directions of the C_γ carbons for those side-chains where C_β is located in the $\mathtt{L}$-α island of Figure 2. We continue to utilize the framing determined by our observer who roller-coasts the C_α backbone with orientation determined by the discrete Frenet frames, and north-pole always in the direction of the next C_α. The result is presented in Figure 4. It reveals that at the level of C_γ, the single $\mathtt{L}$-α island of the C_β becomes divided into two separate but still *highly* localized islands. This reflects the sp3 hybridization of the C_β: There is a putative *gauche*- (*g*-) island where around 70% of the residues in the $\mathtt{L}$-α island are located, and a putative *trans* island for the rest. Interestingly, we do not really see any putative *gauche*+ (*g*+) island.

The amino acid propensities of these two islands is displayed in Figure 5. ASN is the most populous in both C_γ islands. However, the propensity of ASP is elevated

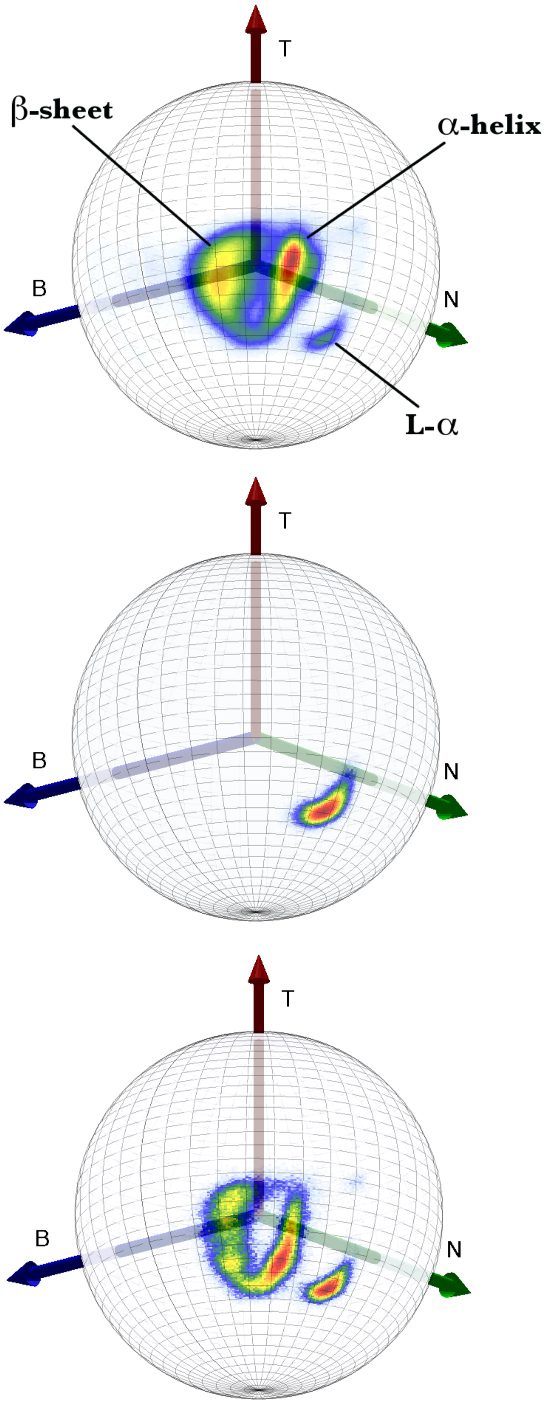

FIG. 2. (Color online:) The directions of the C_β carbons, as seen by our Frenet frame observer who is located at the corresponding C_α carbon which is at the center of the sphere. The vector $\mathbf{t}$ points to the direction of the next C_α carbon. On top all residues in our data set including ASN. In the middle we display only the L-α region of the Ramachandran plot. On bottom we display only those ASN that are in a loop in PDB classification.

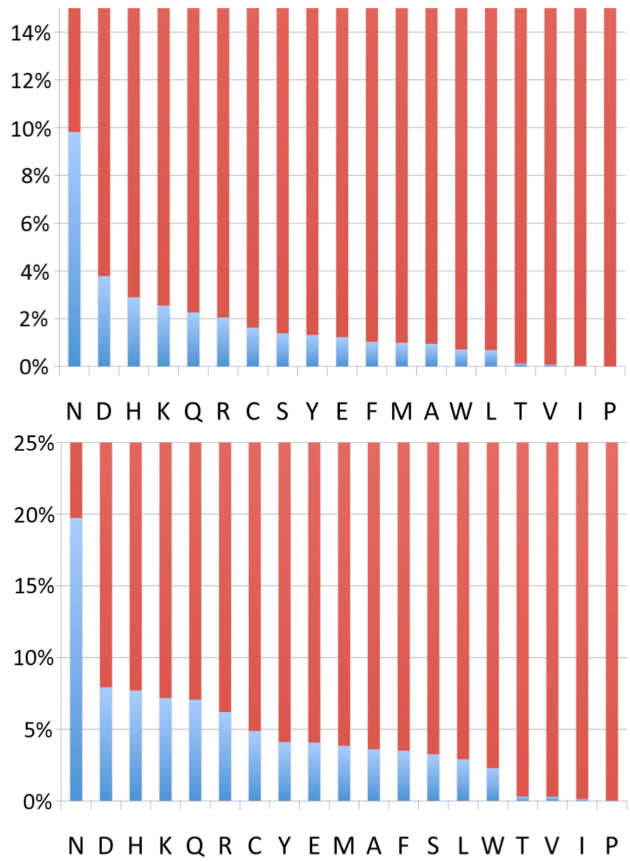

FIG. 3. (Color online:) The percent distribution of non-glycyl residues in the L-α island of Figure 2. In the top Figure we display the result for all amino acids in our entire data set, and in the bottom Figure for those in our data set that are classified as loops in PDB. The propensity of carbonylic ASN (N) is clearly enhanced in both cases. But in both cases the similarly carbonylic ASP (D) has about the same percent-wise propensity with the non-carbonylic HIS (H), and the carbonylic GLU (E) is relatively quite suppressed.

only in *trans* island. In the *g*- island both non-carbonylic HIS (H) and LYS (K) and even the carbonylic GLN (Q) have a higher propensity than ASP. At the moment we have no good explanation for this observation, and we leave it as challenge. In Figure 6 we plot the percentage ratios of the different amino acids as they appear in the two C_γ islands. We note that around 43% of residues in the putative $g-$ island are non-carbonylic, while in the putative *trans* island the number is much lower, close to 12%.

We proceed to the next level along the side-chain, to map the C_δ carbons. in Figure 7 we plot these carbons for those side-chains where C_β is located in the L-α island. In the Figure 7 on top, we show them as they are seen by by our discrete Frenet frame observer who sits at the locations of the C_α carbons. In the Figure 7 on bottom we show them as they are seen in the C$_\beta$ frame for an observer now sitting at the C$_\beta$ location, this time us-

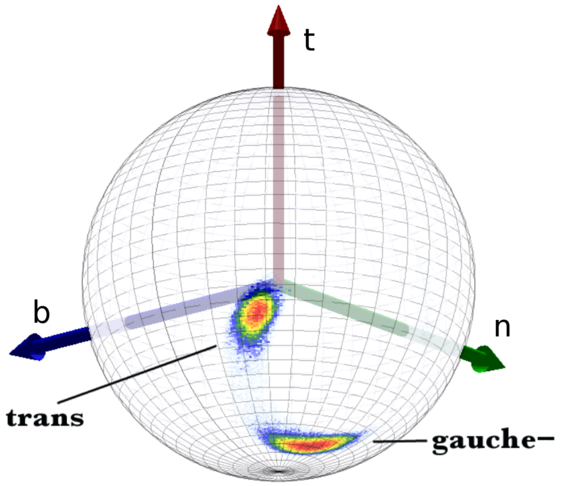

FIG. 4. (Color online:) The directions of those C_γ carbons for which the C_β is located in the L-α island, and as seen by our Frenet frame observer who is located at the center of the C_α carbon which is situated at the center of the sphere.

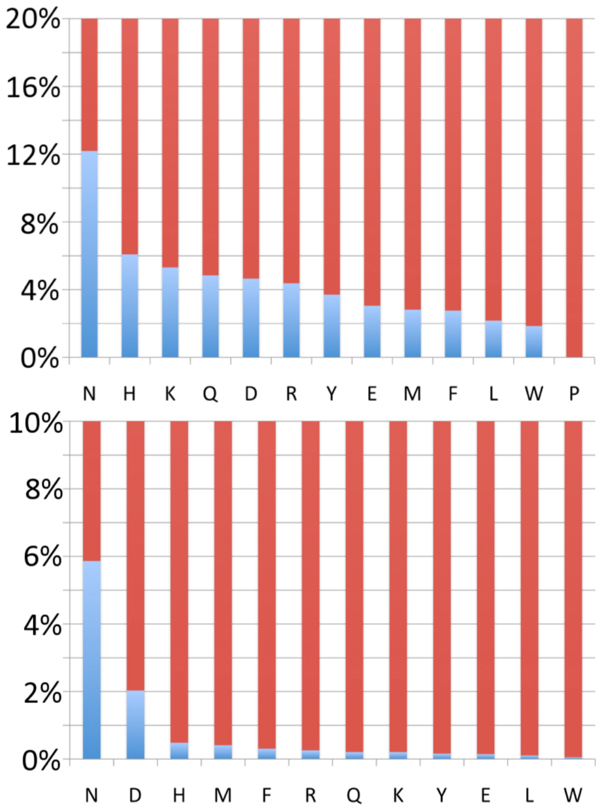

FIG. 5. (Color online:) The percent-wise propensity of different amino acids in the putative *g*- island (top) and *trans* C_γ-island (bottom) in Figure 4

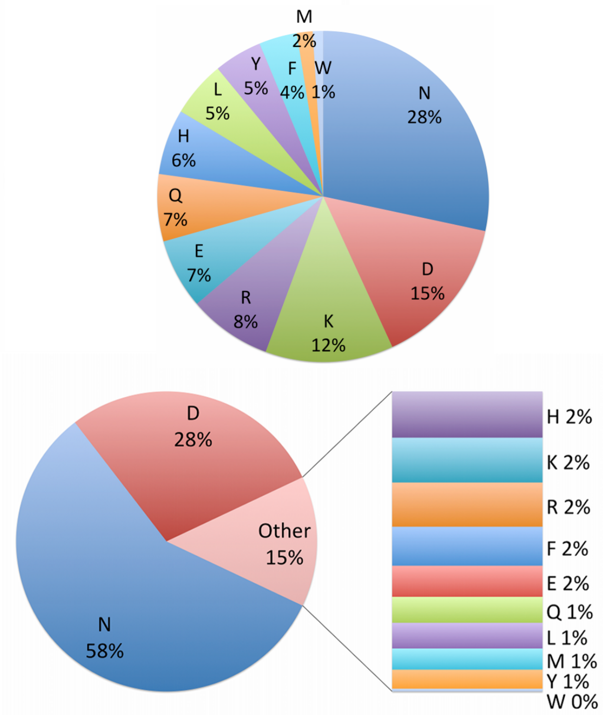

FIG. 6. (Color online:) The relative number of different amino acids in the putative $g-$ (top) and *trans* (bottom) C_γ-islands.

ing the stereographic projection. Since ASN (and ASP) has no C_δ carbon, we display instead the direction of the side-chain O atom for ASN, the result is shown in Figure 8. In the top Figure 8 we use the C$_\alpha$ based Frenet frame observer and in the middle and bottom Figure 8 we use the C$_\beta$ frame observer in combinations with stereographic projection.

From Figure 7 we observe that the directions of the C_δ continue to be highly localized, independently of the type of amino acid. But unlike in the case of C_γ, we find quite surprisingly, that now there is only one clearly visible island. We do not have any definite stereochemistry or physics based explanation why the clearly visible sp3 hybridization based doubling that we observe at the level of C_γ has now completely disappeared. However, we do observe the formation of a second, relatively very weakly occupied island at larger values of the latitude angle and with longitude angle $\chi \sim -2\pi/3$. This island is clearly visible in Figure 7 (right). There is also a third, very faint island in the direction $\chi \sim \pi$ that (barely) becomes visible in the stereographically projected Figure 7 (left). At the moment we do not have a basis to conclude whether the extremely low population of the second and third island is a real effect or only a reflection of problems in the experimental data. We refer to [29], that there are presently an estimated half a million incorrectly positioned side-chain atoms PDB data. In this light, the reason for the sparse population of the two

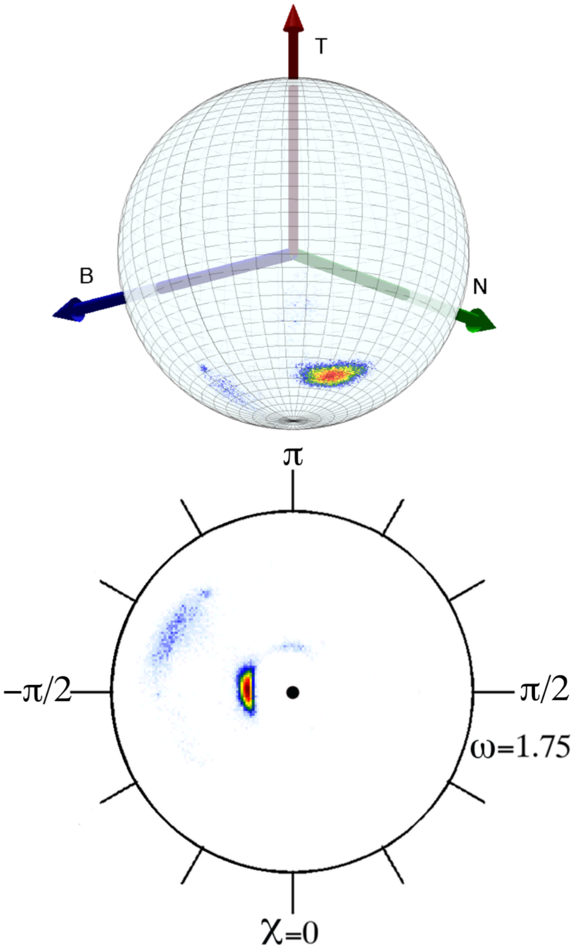

FIG. 7. (Color online:) The directions of the C_δ-carbons in the discrete Frenet frame of the C_α carbons on the surrounding two-sphere (top) and in the C_β frame (bottom). In the Figure on top, the C_β atom is located at the origin of the two-sphere, which is stereographically projected from the north pole (with C_α at the south pole).

additional sp3 hybridized islands should be subjected to experimental curiosity, to determine the cause.

Since ASN has no C_δ carbon, in Figure 8 we display instead the O atoms of the ASN side-chain according to PDB identification. In the top Figure 8 we use discrete Frenet frame of the C_α-carbons, and in the middle and bottom we use the stereographically projected C_β frame. We note that the two C_γ islands appear to become divided into four distinct but still highly localized islands. However, we recall that the identification between the ASN side-chain O and N can be very difficult, and there are apparently numerous errors in the O and N identifications in PDB data [29]. Thus we have displayed in Figure 8 (top) the N atoms according to PDB identification as well. By comparing the Figures 8 (middle) and (bottom) we propose that *most likely* the two inner-most islands denoted **a** and **b** in Figure 8 describe N instead of

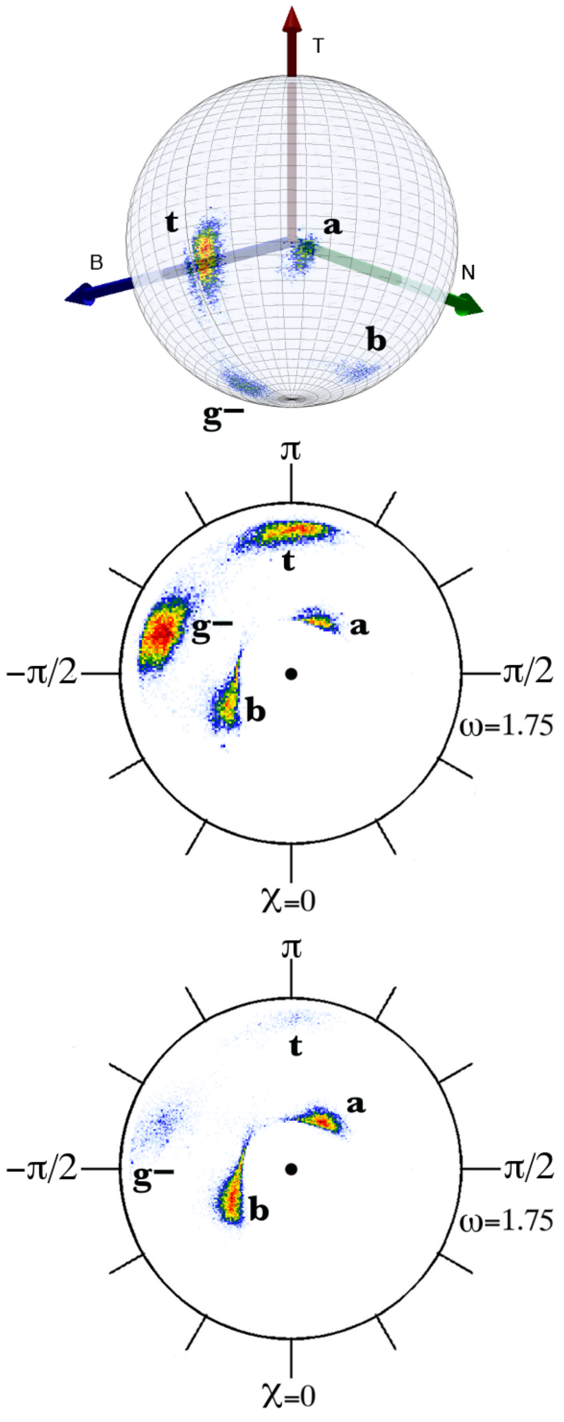

FIG. 8. (Color online:) On top, the directions of the side-chain O atoms of ASN in the discrete Frenet frame of the C_α carbons. In the middle and bottom we use the same stereographically projected C_β frames as in Figure 7 (right). The Figure in middle displays the same atoms as the Figure on top. On bottom, the directions of the side-chain N atoms of ASN suggests that the correct identification of **a** and **b** regions in the top and middle figure should be N and not O as in PDB. Similarly, the regions **t** and **g-** in the bottom should be O and not N in PDB. See [29]. (**t** is *trans* and **g-** is *gauche-*)

O atoms. If so, our visualization technique could become a useful tool in detecting erroneously identified O and N atoms and help to resolve the kind of issues raised in [29]. This could be scrutinized by a careful re-analysis of high resolution x-ray crystallography data.

Finally, in Figure 9 we have mapped the locations of the C_γ atoms in our entire dataset (except for prolines) as they are seen in the stereo graphically projected $C-\beta$ frame. In the Figure at top we show all atoms except those that have C_β in the L-α, and in Figure at bottom we show only the L-α atoms. The sp3 hybridization of the C_β covalent bond structure is clearly visible. Furthermore, in each of the three regions in left we recognize the substructure that correspond to the α-helices, β-strands and the interconnecting loops. Each of the three regions is then isomorphic to Figure 2a.

Obviously, it is straightforward to continue the present analysis to inspect additional side-chain atoms. However, here our goal is not to perform a detailed and complete analysis of all the side-chain atoms, we simply aim to describe a method.

V. SOLITONS

The localization we have observed in the L-α side-chain atoms proposes that there is an organizational principle in the side-chain orientations that extends beyond a single peptide unit. Hence it can not be detected by the Ramachandran plot or in terms of the standard rotamer libraries, these only provide information on a given peptide unit. The backbone C_α atoms we have inspected all correspond to the L-α position of the C_β, this region is known to commonly appear in connection of loops in lieu of regular secondary structures. Thus the order we have observed is *a priori* not a reflection of any apparently regular secondary structure category at the level of the backbone geometry, but a characteristic of loops. We propose that it is due to the presence of a soliton solution to a discrete version of nonlinear Schrödinger (DNLS) equation that universally describes the backbone C_α geometry in (practically) all folded proteins.

For the soliton description we do not need to know the atomic level details of the energy function. We only need to apply general symmetry principles to the abstract full quantum mechanical, all-atom Hamiltonian operator $\mathcal{H}[q_i, p_i]$. Here the index i = 1, ..., N labels all pairs of canonical coordinates (q_i, p_i) that describe the elementary constituents. These include the individual C, O, N, H and every other atom in the protein and in the solvent. We also account for the valence electrons, and for every local and long range interaction between all the atoms both in the protein and in the solvent. For simplicity we take all the variables to be point-like, that is we work at the first quantized level. The canonical partition function is computed by the path integral,

$$\mathcal{Z} = Tre^{-\beta\mathcal{H}} = \int [dq] e^{-\frac{1}{\hbar}\mathcal{S}(q,\dot{q})}$$

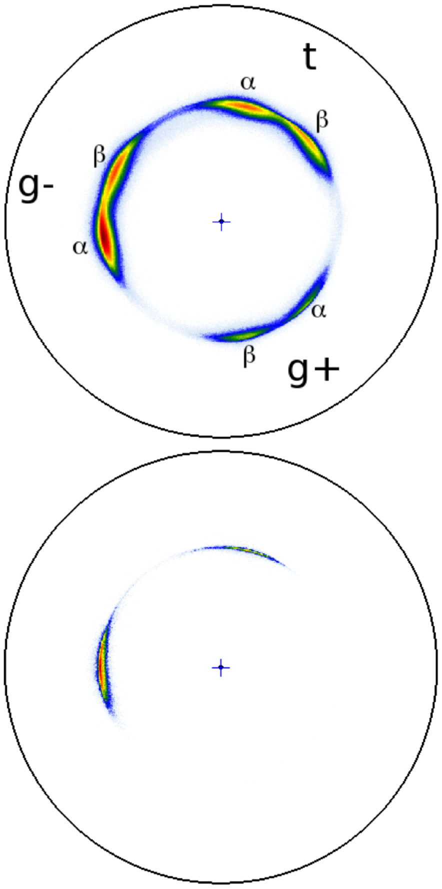

FIG. 9. (Color online:) The directions of the C_γ atoms in the C_β centered frames, on two-sphere stereographically projected from its north pole and with C_α at south-pole. On the top all those C_γ atoms for which the C_β is not in the L-α, and on bottom those that correspond to C_β in the L-α island. (*gauche-* on left, **t** on top-right and *gauche+* on bottom-right.)

where $\mathcal{S}(q, \dot{q})$ is the classical Euclidean action of $\mathcal{H}[q_i, p_i]$. The integration extends over all period configurations (anti-periodic in the case of fermions). With the partition function we obtain the thermodynamical Helmholtz free energy as follows: We introduce external sources $j_i(t)$ and extend the partition function into the generating functional of connected Green's functions,

$$W[j] = \ln \left\{ \int [dq] e^{-\frac{1}{\hbar}\mathcal{S}_\beta + \frac{1}{\hbar}\int\limits_0^\hbar \beta j \dot{q}} \right\}$$

We introduce the Legendre transformation

$$\Gamma[q] = -\frac{1}{\beta}\left\{ W[j] - \int\limits_0^{\hbar\beta} \frac{\delta W}{\delta j} \cdot j \right\}$$

This defines the effective action that coincides with the Helmholtz free energy when we take the limit $j_i \to 0$.

There are various methods to compute the Helmholtz free energy E. Here we introduce a finite difference version of the gradient expansion. In the leading nontrivial order we have

$$E = \lim_{j_i \to 0} \Gamma[q] = \sum_i V^{(i)}(q_i) + F^{(i)}(q_i, q_{i+1}) + \dots$$

$$= \sum_i V^{(i)}(q_i) + \frac{\partial F^{(i)}(q_i, q_{i+1})}{\partial q_{i+1 | q_{i+1}=0}} q_{i+1} \ + \dots \qquad (7)$$

The potentials $V^{(i)}$ are all local and the $F^{(i)}$ are bi-local. The higher order terms in the expansion are either higher order polynomials in the nearest neighbor variables, or terms that introduce couplings between next-to-nearest neighbor variables.

In the expansion (7), in the case of the backbone we identify the generalized coordinates q_i with the bond and torsion angles (κ_i, τ_i) in (2), (3). We assume that all the additional variables that appear in the full Hamiltonian operator $\mathcal{H}[q_i, p_i]$ have been integrated over in constructing the partition function. They affect the detailed functional form of the coefficients $V^{(i)}$, $F^{(i)}$ *etc.* in (7). Since (5) contains only the $\mathbf{t}_i$, it is clear that the expansion (7) in terms of (κ_i, τ_i) must remain invariant if we introduce a local frame rotation in the normal plane spanned by $(\mathbf{n}_i, \mathbf{b}_i)$. This introduces a strong constraint to its functional form. In [30], [31] this has been utilized to show that in the leading order the expansion (7) is uniquely determined. It can only contain the following terms [6]-[8], [31],

$$E = -\sum_{i=1}^{N-1} 2\,\kappa_{i+1}\kappa_i + \sum_{i=1}^{N} \bigl\{ 2\kappa_i^2 + q \cdot (\kappa_i^2 - m^2)^2$$

$$+ \frac{d_\tau}{2}\,\kappa_i^2 \tau_i^2 - b_\tau \kappa_i^2 \tau_i - a_\tau \tau_i + \frac{c_\tau}{2} \tau_i^2 \biggr\} \qquad (8)$$

Here the first sum together with the three first terms in the second sum coincide with the integrable energy function of the conventional DNLS equation with a potential that displays spontaneous symmetry breaking. The fourth (b_τ) and the fifth (a_τ) terms are the *only* two lower order nontrivial conserved quantities that appear in the integrable DNLS hierarchy prior to the energy. These are the momentum and the helicity, respectively. The last (c_τ) term is the standard Proca mass term. The parameters are all global and specific only to a super-secondary structure such as helix-loop-helix. In particular they are independent of the detailed nature of amino acids.

Unlike a force field in molecular dynamics, the energy function (8) does not describe the fine details of the atomary level interactions such as Coulomb, van der Waals, hydrogen bonding *etc.* Instead, like an effective Landau-Lifschitz theory it describes the properties of a folded protein backbone in terms of universal physical arguments. Full details and motivation of (8) are presented in [6]-[8], [31]

The remarkable property of (8) is that the torsion angle τ_i is only subject to local interactions, all explicit non-local interactions are carried by the bond angle κ_i. Furthermore, since τ_i appears at most quadratically we can solve for it in terms of κ_i,

$$\tau_i = \frac{a_\tau + b_\tau \kappa_i^2}{c_\tau + d_\tau \kappa_i^2} \qquad (9)$$

When we substitute this into the variational equation of κ_i that follows from (8), we arrive at a generalized version of the DNLS equation. Its soliton solution has been constructed in [6]-[8]. In particular, it has been observed that the soliton can be approximated by the discretized version of the soliton solution of the continuum dark NLSE soliton [9]-[11],

$$\kappa_i = \frac{m_1 \cdot e^{c_1(i-s)} - m_2 \cdot e^{-c_2(i-s)}}{e^{c_1(i-s)} + e^{-c_2(i-s)}} \qquad (10)$$

Here the various parameters each have a natural interpretations, see [6]-[8] for a detailed description: The parameter s determines the backbone site location of the center of the fundamental loop that is described by the soliton. The values of the parameters $m_{1,2} \in [0, \pi]$ $mod(2\pi)$ are entirely determined by the bond angles of the adjacent helices and strands. Finally, *only* the c_1 and c_2 are intrinsically loop specific parameters, they specify the length of the loop. The soliton profile of κ_i determines the torsion angles τ_i by (9). According to [8] practically all PDB proteins can be constructed as the sum of terms of the form (10), in a modular fashion from a relatively small number of soliton profiles.

Following [28] we argue that in the Frenet frames, the angular positions of the side-chain atoms can be similarly determined in terms of the corresponding κ_i values only. For this we denote by (θ, ϕ) the standard spherical latitude and longitude angles of the sphere that surrounds the C$_\alpha$ observer. We propose that to leading order in the expansion (7), in these coordinates each of the side-chain atoms has an energy function that has the same functional form as the energy function of the backbone torsion angles. Consequently, for each side-chain atom we introduce only the following two leading contributions to the energy

$$E_\theta = \sum_{i=1}^{N} \left\{ \frac{d_\theta}{2}\,\kappa_i^2 \theta_i^2 - b_\theta \kappa_i^2 \theta_i - a_\theta \theta_i + \frac{c_\theta}{2} \theta_i^2 \right\} \qquad (11)$$

$$E_\varphi = \sum_{i=1}^{N} \left\{ \frac{d_\varphi}{2}\,\kappa_i^2 \varphi_i^2 - b_\varphi \kappa_i^2 \varphi_i - a_\varphi \varphi_i + \frac{c_\varphi}{2} \varphi_i^2 \right\} \qquad (12)$$

Note that these contributions have been *carefully selected* so that they will *not* change the functional form of(8). The addition of (11), (12) will only redefine the coefficients in the κ_i dependent terms in (8) which does not lead to any change in the underlying soliton structure. In particular, we can still utilize the approximative soliton profile (10). In parallel with (9) the spherical angles (θ_i, φ_i) for each of the side-chain atoms are then dynamically determined by the DNLS soliton profile of the backbone bond angles κ_i,

$$\theta_i = \frac{a_\theta + b_\theta \kappa_i^2}{c_\theta + d_\theta \kappa_i^2}$$

$$\varphi_i = \frac{a_\varphi + b_\varphi \kappa_i^2}{c_\varphi + d_\varphi \kappa_i^2}$$

The present visual analysis implies that in the case of a L-α residue the numerical values of both (b_θ, d_θ) and (b_φ, d_φ) are vanishingly small for the C$_\beta$, C$_\gamma$ and C$_\delta$ carbons, and for the side-chain N and O atoms in the case of ASN and ASP. But both (a_θ, c_θ) and (a_φ, c_φ) have amino acid independent, finite and *universal* values that can be directly inferred from the Figure 2 (middle), 4, 7 and 8 respectively,

$$< \theta_i >_{\mathtt{L}-\alpha} \approx \frac{a_\theta}{c_\theta}$$

$$< \varphi_i >_{\mathtt{L}-\alpha} \approx \frac{a_\varphi}{c_\varphi}$$

We now proceed to argue that these universal values can be understood in terms of relatively few DNLS solitons. We then show how these solitons can be classified using our graphical tools.

VI. SOLITON VISUALIZATION

It has been argued in the literature that in the case of ASN and ASP the L-α Ramachandran region become stabilized by a local but non-covalent attractive interaction between the side-chain and backbone carbonyls, with the backbone oxygen atom in a special rôle [13], [15]. Unlike the Ramachandran plot, our Frenet framing can provide information on the neighboring peptide units and we have investigated the directions of all backbone O atoms in our data set, in a group of residues around the i^{th} side-chain C_β that is located in the L-α island. The result shown in Figure 10 displays how these O atoms are seen by our Frenet frame observer who is located at the i^{th} central C_α carbon. Our observer finds that the directions of the nearby backbone O atoms are very srongly localized and correlated. The localization is residue *independent* and extends itself over at least four different residues:

- For the $i-2$ site there is strong localization with a three-fold degeneracy that is reminiscent of the *trans/gauche* (sp3 hybridization) degeneracy. The data is consistent with vanishing values of both (b_θ, d_θ) and (b_φ, d_φ).
- For the site $i-1$ we have very strong localization along the longitudinal (φ_{i-1}) direction, with a tiny oscillation in the latitudinal (θ_{i-1}) direction. For the corresponding energy, (b_φ, d_φ) are again vanishingly small while (b_θ, d_θ) are now small but non-vanishing.
- For the site i we have a single localized oscillator in the longitudinal direction. Thus (b_φ, d_φ) are now small but non-vanishing while (b_θ, d_θ) vanish.
- For the site $i+1$ we again find the three-fold *trans/gauche* degeneracy: There are three oscillators in the longitudinal direction, and they are all located very close to the north-pole. Consequently (b_θ, d_θ) vanish while (b_φ, d_φ) do not. In fact, the φ_{i+1} amplitudes are quite large.

The localization pattern of the backbone O atoms means that for our Frenet frame observer the backbone geometry around a L-α residue shows very little variations. Only a very limited set of extended backbone geometries are accessible. Since the regime that covers the sites from the $(i-2)^{th}$ to the $(i+1)^{th}$ involves four sets of bond and torsion angles, each of them defined in terms of three *resp.* four residues we conclude that the geometries reflect the non-local collective interplay of *at least up to seven* different residue sites along the backbone. This is in line with our proposal that the positions of the side-chain atoms are determined dynamically by the backbone, in terms of a small number of different DNLS soliton profiles according to (11), (12).

To expose the soliton structures that surround the L-α island, we consider the distribution of the backbone bond and torsion angles that are attached to those C_α carbons where the C_β atom is in the L-α position. The result is shown in Figure 11 on a stereographically projected two-sphere, separately for ASN and ASP and for the remaining non-glycyl amino acids.

We observe no practical difference between the residues. Nor do we find any practical difference between the various *trans* and *g-* positions. Instead, we do observe the following general pattern: For the backbone C_α-C_α link that precedes the L-α island, three different regions on the (κ, τ) plane are probable. These are the regions that we have denoted with **a**, **b** and **c** respectively in the Figure 12 (top); In this Figure we have combined all the data that are displayed separately in the parts **a, b, c** and **d** of Figure 11. After the L-α island there are also three different regions that are probable. We denote these regions with letters **b** and **d** and **e** respectively in the Figure 12 (bottom), now combining the data in parts **e, f, g, h** in Figure 11. Note that the regions **b** in the two parts of Figure 12 practically overlap.

By inspecting the protein structures in our data set we conclude that the presence of a residue in the L-α island causes the following phenomenological *selection rules* between the regions displayed in Figure 12. When we roller-coast along the backbone:

- The region **a** can only precede regions **d** and **e**.
- Both regions **b** and **c** can be followed by any of the

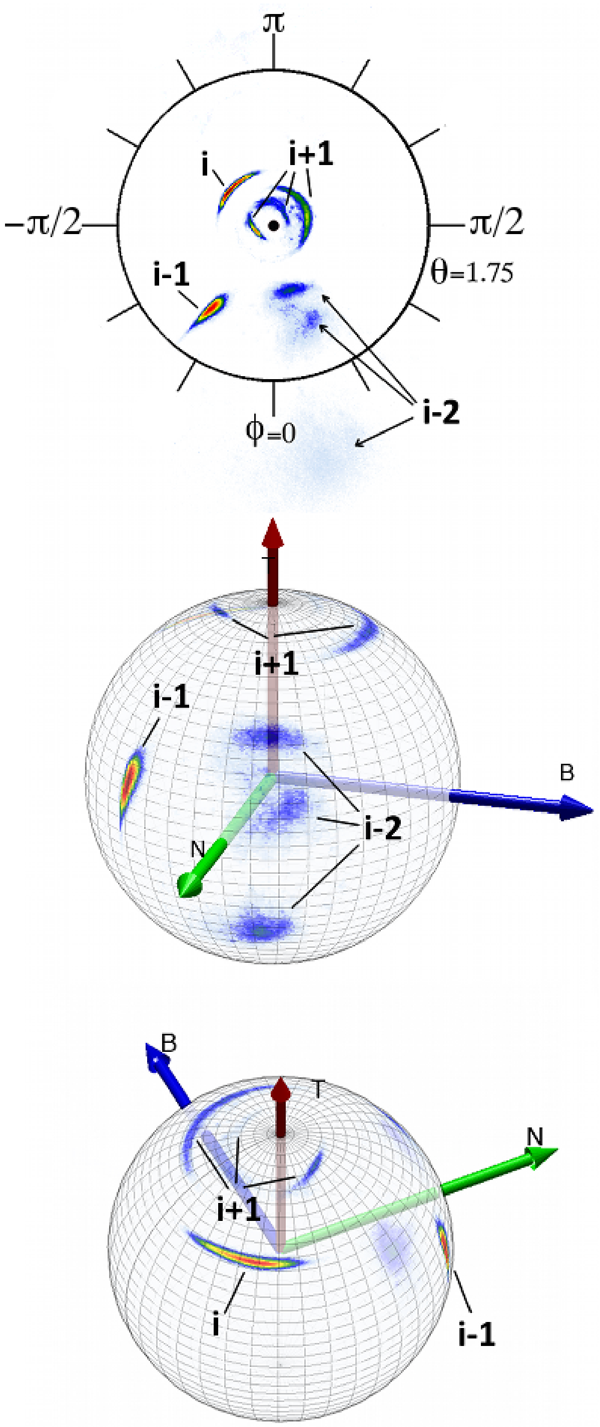

FIG. 10. (Color online:) The orientations of backbone O atoms around the site i that is located in the L-α island, as seen by our discrete Frenet frame observer at the i^{th} central C_α-carbon, on a stereographically projected two-sphere (top) and on the surrounding two-sphere that we have displayed from two different perspectives (middle, bottom). For the i^{th} and $(i-1)^{th}$ atom only one position appears to be available while the $(i+1)^{th}$ and $(1-2)^{th}$ atoms each have three available (*trans/gauche*) positions. The angle ϕ is measured from the **N** axis.

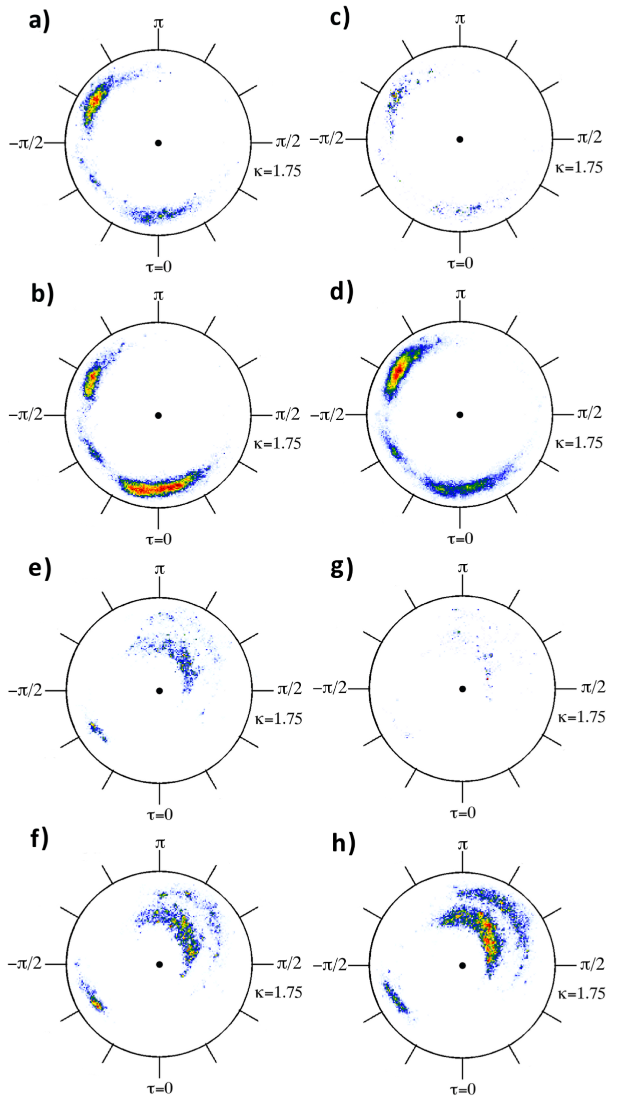

FIG. 11. (Color online:) The (κ, τ) distributions for backbone links that are attached to a C_α carbon with C_β in the L-α island on stereographically projected two-sphere as in Figure 1. We display separately ASN, ASP, and all the rest. On left column the C_α carbons in the case where the corresponding C_γ carbon is in the *trans* island, on right for those where the C_γ is in the *g-* island. First row a),c) is for link that precedes either ASN or ASP. Second row b),d) is for link preceding any other non-glycyl amino acid. Third row e),g) is for link following either ASN or ASP. Fourth row f),h) is for link following the others.

three regions **b**, **d** and **e**.

- The residue preceding either **a** or **c** is not located in the L-α island.

- Both the residue preceding and following **b** can be located in the L-α island.

- If the two residues following **c** are both in the L-α island, the first residue connects **c** to **b** and the second connects **b** to **b**.

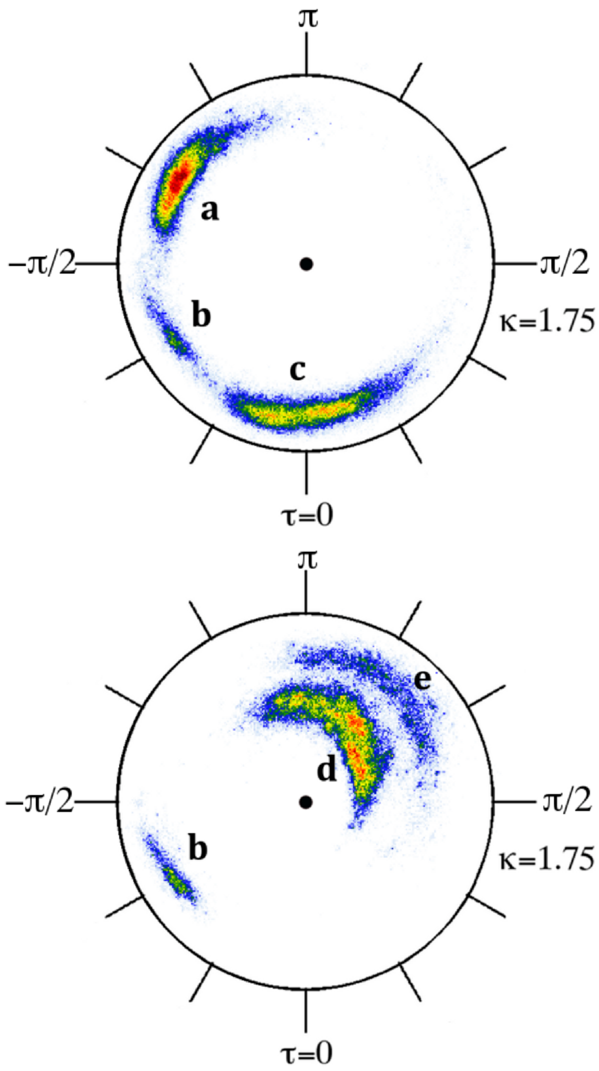

FIG. 12. (Color online:) The (κ, τ) distributions for all backbone links that are attached to a residue in the L-α island. On the top for the link preceding the residue in the L-α island, on the bottom following the residue in the L-α island.

These are the selection rules, that limit the available global topology of the backbone solitons when we pass the L-α site. Notice that since the region **b** has the same bond angle as the standard α-helix region and since the torsion angles are equal in magnitude but have an opposite sign, a repeated structure in **b** is the right-handed mirror image of the standard α-helix. Consequently this is truly the region of *left-handed α-helices.*

These selection rules classify the different possible DNLS soliton profiles in the presence of the L-α island. In particular, we have found that there seems to be no more than four solitons that are particularly common around a residue with C_β located in the L-α island. We now describe these solitons qualitatively using our visual tools, as transition trajectories between the different regions that appear in the maps of Figure 1. These transitions illustrate how our observer turns at the location of each C_α as she roller-coasts through the soliton. The results are shown in Figure 13. In each case the pink arrow corresponds to a site where C_β is located in the L-α island; Recall that the bond and torsion angles are link variables, they connect two C_α carbons according to (4). Clearly, the Figure 13 is but an example of a general method to visually classify solitons in a manner that directly reflects the geometry of the underlying backbone.

- In Figure 13a, the first residue takes our observer away from an α-helix region to the region **a** in the Figure 12 left (black arrow). This is followed by a residue in the L-α island, that takes the observer to the region **d** in the Figure 12 right (pink arrow). Finally, there is a transition to the β-strand region (black arrow). Consequently this is a short soliton that takes us from the ground state which is an α-helix to the other ground state which is a β-strand.

- The second soliton trajectory shown in Figure 13b starts from the β-strand region with a residue that takes the observer into region **c** in Figure 12 left. The following residue that is located in the L-α island then causes a transition into region **d** in Figure 12 right (pink arrow). This is followed by a transition back to a β-strand region. Since the initial and final positions are in a β-strand, this is an example of a soliton that combines the β-strand with another β-strand.

- The third trajectory that we have described in Figure 13c starts from the β-strand region and proceeds to region **c** in Figure 12 left. From there the trajectory proceeds to region **b** in Figure 12 left, with the transition caused by a residue in the L-α island. This is followed by a transition to region **d** and then back to the β-strand region. This trajectory is also an example of a soliton that combines the β-strand with another β-strand.

- Finally, the fourth trajectory that is also common in our data set is the one displayed in Figure 13d. It is similar with the trajectory described in Figure 13c, except that now the residue that is located in L-α island causes the transition from **b** to **d** in Figure 12. This trajectory is also an example of a soliton that combines the β-strand with another β-strand.

The remarkable property of solitons c) and d) in Figure 13 is, that they have similar overall topology and differ from each other only by the location of the L-α along the trajectory. It is quite plausible that in some proteins these two solitons are but two states of an oscillating discrete "breather" soliton. The ensuing proteins are presumable unstructured.

Finally, for the purposes of soliton taxonomy we note that when we analyze the proteins in the version v3.3 Library of chopped PDB files for representative CATH domains we find that the propensity of our solitons is largest in the (mainly-β) CA level classes 2.90, 2.160 where over 5% of all residues are in the L-α island. We also find that any CA level family has at least 1% of their residues in the island, except 1.40 where the single representative

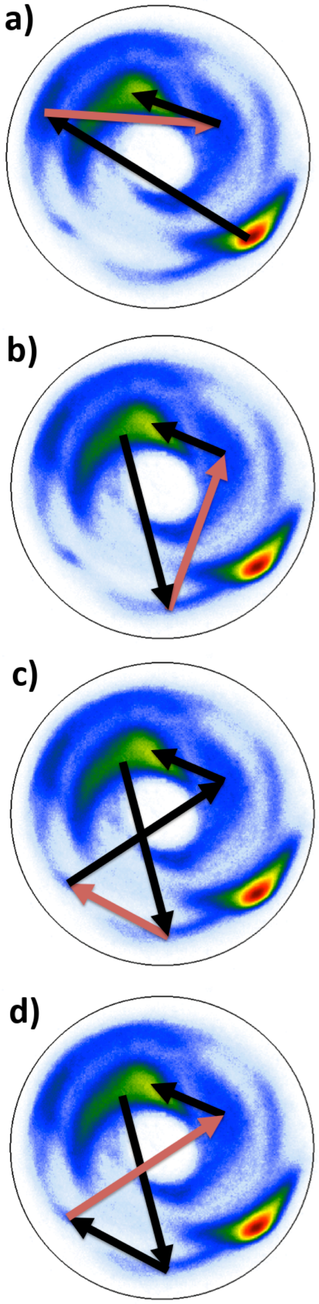

FIG. 13. (Color online:) Four different soliton trajectories through a residue in the L-α island that are common in our data set, on the stereographically projected two-sphere. The arrow shows how the Frenet frame observed sees the soliton to proceed from a C_α to the next C_α. In each case the pink line denotes the transition that is caused by the presence of a residue in the L-α island. The trajectory **a** described a soliton that connects the α-helix region to the β-strand region. The remaining ones all both start and end in the β-strand region.

with PDB code 1PPR has no residues in the island.

VII. CONCLUSION

We have developed a new visualization method of proteins. In the case of backbones our method provides information about the geometry of neighboring peptide units. This enables us to go beyond the regime of the canonical Ramachandran plot which does not contain information on the neighboring units. As an example of side chains, we have visually investigated the non-glycyl residues that are located in the L-α region of the Ramachandran plot. *Independently* of the amino acid, we find that for a discrete Frenet frame observer who roller-coasts along the backbone C_α carbons the corresponding side-chain C_β carbons are always localized in the same direction which is clearly different from the direction of the C_β carbons in the right-handed region. This universality in the orientation persists when we investigate the C_γ and C_δ carbons, and the side chain O and N atom in the case of ASN and ASP. The results suggest that instead of reflecting *only* a local interaction between a given backbone unit and its residue, the L-α island is also associated with a largely residue independent backbone conformation.

When we proceed to analyze the distribution of those backbone bond and torsion angles that are associated with the links that both precede and follow a residue that is located in the L-α island, we find that independently of the residue these angles display very similar patterns. Since the definition of a bond angle takes three C_α carbons and the definition of a torsion angle takes four, this prompts us to propose that the geometrical structure associated with the presence of a residue in the L-α island is a soliton that reflects the interplay of at least seven consecutive backbone units. In particular, we have not been able to pin-point any obvious local reason (charged, polar, acidic, hydrophobic/philic) to explain the presence or absence of a residue on the L-α region.

Our approach is based on a novel visualization method to depict proteins. This method is based on advances in three dimensional visualization techniques that have been developed after Ramachandran presented his plot. In the course of our analysis we have been able to observe several systematic patterns including potential anomalies in the PDB data. The visualization method we have developed shows promise to become a valuable tool for both experimental and theoretical protein structure analysis and fold description, in particular for visually describing and classifying the backbone solitons and as a complement to existing side-chain rotamer libraries.

ACKNOWLEDGEMENT

We thank S. Hu and J. Åqvist for discussions.

[1] G. Ramachandran, C. Ramakrishnan and V. Sasisekharan, Journ. Mol. Biol. **7** 95 (1963)

[2] C. Ramakrishnan and G. Ramachandran, Biophys. J. **5** 909 (1965)

[3] J. Janin, S. Wodak, M. Levitt, D. Maigret, J Mol Biol. **125** 357 (1978)

[4] A.J. Hanson, *Visualizing Quaternions*, Morgan Kaufmann Elsevier (London, 2006)

[5] J.B. Kuipers, *Quaternions and Rotation Sequences: a Primer with Applications to Orbits, Aerospace, and Virtual Reality*, Princeton University Press (Princeton, 1999)

[6] M.N.Chernodub, S. Hu and A.J. Niemi Phys. Rev. **E82** 011916 (2010)

[7] N.Molkenthin, S. Hu and A.J. Niemi Phys. Rev. Lett. **106** 078102 (2011)

[8] A. Krokhotin, A.J. Niemi and X. Peng, arXiv:1109.3903v1 [physics.bio-ph]

[9] A.S. Davydov, Journ. Theor. Biol. **66** 379 (1977)

[10] P.G. Kevrekidis, *The Discrete Nonlinear Schrdinger Equation: Mathematical Analysis, Numerical Computations and Physical Perspectives* (Springer-Verlag, Berlin, 2009)

[11] L.D. Faddeev and L.A. Takhtajan, *Hamiltonian methods in the theory of solitons* (Springer Verlag, Berlin, 1987)

[12] S.Hu, M. Lundgren and A.J. Niemi Phys. Rev. **E83** 061908 (2011)

[13] C.M. Deane, F.H. Allen, R. Taylor and T.L. Blundell, Prot. Eng. **12** 1025 (1999)

[14] F.H. Allen, C.A. Baalham, J.O.M. Lommerse and P.R. Raithby, Acta Crystallog. **B54** 320 (1998)

[15] P. Chakrabarti and D. Pal, Prog. Biophys. Molec. Biol. **76** 1 (2001)

[16] N.E. Robinson and A.B. Robinson, PNAS (USA) **98** 12409 (2001)

[17] C.R. McCuddena and V.B. Kraus, Clinic. Biochem. **39** 1112 (2006)

[18] N.E. Robinson and A.B. Robinson, *Molecular Clocks Deamidation of Asparaginyl and Glutaminyl Residues in Peptides and Proteins* Althouse Press (London, 2004)

[19] E.H. Koo, P.T. Lansbury and J.W. Kelly, PNAS (USA) **96** 9989 (1999)

[20] R.L. Bishop, Am. Math. Monthly **82** 246-251 (1974)

[21] R.S. Hartenberg and J. Denavit, *Kinematic synthesis of linkages* McGraw-Hill (New York, NY, 1964)

[22] H.M. Berman, K. Henrick, H. Nakamura and J.L. Markley, *Nucl. Acids Res.* **35** (Database issue) D301 (2007)

[23] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells and J.M. Thornton, Structure **5** 1093 (1997)

[24] L. Schäfer and L.M. Cao, Journ. Mol. Struc. **333** 201 (1995)

[25] P.A. Karplus, Prot. Sci. **5** 1406 (1996)

[26] D.S. Berkholz, M.V. Shapovalov, R.L. Dunbrack, Jr. and P.A. Karplus, Structure **17** 1316 (2009)

[27] W.G. Touw and G. Vriend, Acta Cryst. **D66** 1341 (2010)

[28] M. Lundgren and A.J. Niemi, arXiv:1109.0423v1 [q-bio.BM]

[29] C.X. Weichenberger and M.J. Sippl, Nucl. Acids Res. **35** (Web Server Issue) W403 (2007)

[30] A.J. Niemi, Phys. Rev. **D67** 106004 (2003)

[31] U.H. Danielsson, M Lundgren and A.J. Niemi, Phys. Rev. **E82** 021910 (2010)